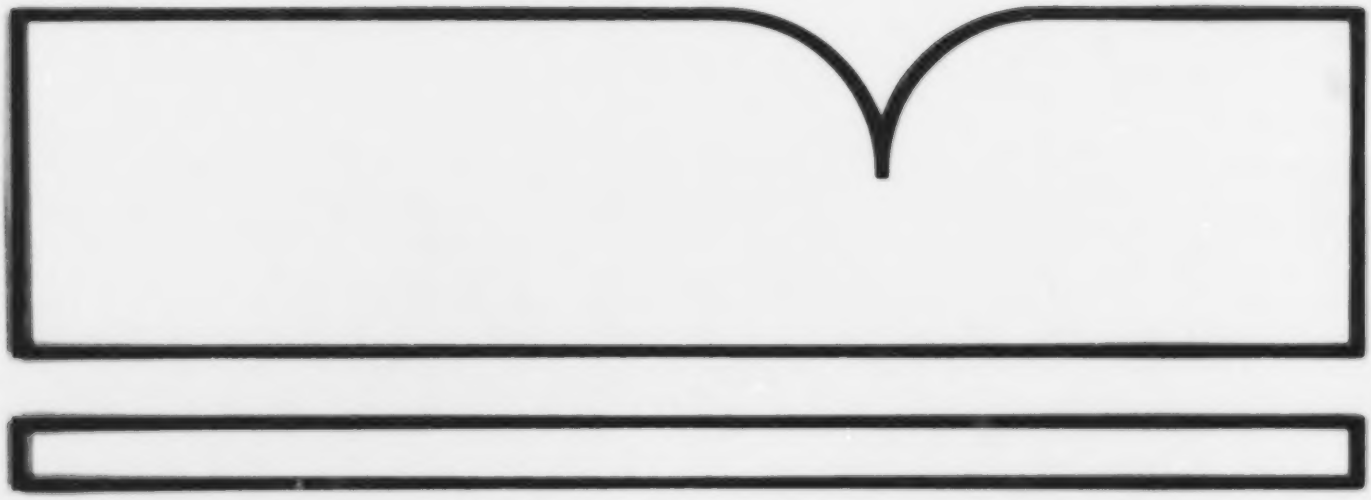
PB87-115333

Analysis of the Volume of Red Blood Cells Application of the Expectation-Maximization Algorithm to Grouped Data from the Doubly-Truncated Lognormal Distribution

(U.S.) Health Effects Research Lab.
Research Triangle Park, NC

1986



U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before comple

1. REPORT NO. EPA/600/J-86/172	2.	3. PB87-115333
4. TITLE AND SUBTITLE Analysis of the Volume of Red Blood Cells: Application of the Expectation-Maximization Algorithm to Grouped Data from the Doubly-Truncated Lognormal Distribution (Journal Version)		5. REPORT DATE 1986 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) C.E. McLaren and G. M. Brittenham V. Hasselblad		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Case Western Reserve University U.S. EPA, HERL, Biometry Division, MD-55		10. PROGRAM ELEMENT NO. 11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Health Effects Research Laboratory - RTP, NC Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, NC 27711		13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE EPA/600/11

15. SUPPLEMENTARY NOTES

REF: Biometrics, 42:143-158, March 1986

16. ABSTRACT

In accordance with general principles recommended by the International Committee for Standardization in Haematology (1982, *Journal of Clinical Pathology* 35, 1320–1322), we have developed statistical methods for the analysis of red cell volume distributions. To select an appropriate reference distribution for goodness-of-fit testing, we derived a mathematical model of erythropoiesis that predicted a lognormal form for the distribution of erythrocyte volumes. Model predictions were then tested using samples obtained from 50 healthy individuals. Each grouped red cell volume distribution was doubly-truncated to eliminate artifactual frequency counts. Distribution parameter estimates were computed using the expectation–maximization algorithm, a missing information technique. Results of the one-sample chi-square goodness-of-fit test showed a fairly even distribution of *P*-values over the interval [0, 1]. Examples of the application of these statistical procedures to distributions from patients with anemia are given. Our results suggest that, for the analysis of red blood cell volumes, (i) parameter estimation should be made with the expectation–maximization method, and (ii) the truncated lognormal distribution should be used as a reference distribution for goodness-of-fit testing. This method could be applied to any set of grouped doubly-truncated data which, after transformation, follows the normal model.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	18
	20. SECURITY CLASS (This page) UNCLASSIFIED	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

EPA/600/J-86/172
JOURNAL ARTICLE

BIOMETRICS 42, 143–158
March 1986

NTIS is authorized to reproduce and sell this report. Permission for further reproduction must be obtained from the copyright owner.

Analysis of the Volume of Red Blood Cells: Application of the Expectation–Maximization Algorithm to Grouped Data from the Doubly-Truncated Lognormal Distribution

Christine E. McLaren

Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, and Department of Medicine, Case Western Reserve University at Cleveland Metropolitan General Hospital, Cleveland, Ohio 44106, U.S.A.

Gary M. Brittenham

Department of Medicine, Case Western Reserve University at Cleveland Metropolitan General Hospital, Cleveland, Ohio 44106, U.S.A.

and

Victor Hasselblad

Center for Health Policy Research and Education, Duke University, Durham, North Carolina 27706, U.S.A.

SUMMARY

In accordance with general principles recommended by the International Committee for Standardization in Haematology (1982, *Journal of Clinical Pathology* **35**, 1320–1322), we have developed statistical methods for the analysis of red cell volume distributions. To select an appropriate reference distribution for goodness-of-fit testing, we derived a mathematical model of erythropoiesis that predicted a lognormal form for the distribution of erythrocyte volumes. Model predictions were then tested using samples obtained from 50 healthy individuals. Each grouped red cell volume distribution was doubly-truncated to eliminate artifactual frequency counts. Distribution parameter estimates were computed using the expectation–maximization algorithm, a missing information technique. Results of the one-sample chi-square goodness-of-fit test showed a fairly even distribution of *P*-values over the interval [0, 1]. Examples of the application of these statistical procedures to distributions from patients with anemia are given. Our results suggest that, for the analysis of red blood cell volumes, (i) parameter estimation should be made with the expectation–maximization method, and (ii) the truncated lognormal distribution should be used as a reference distribution for goodness-of-fit testing. This method could be applied to any set of grouped doubly-truncated data which, after transformation, follows the normal model.

1. Introduction

The first and still most widely used method for the diagnostic assessment of changes in red blood cell size distributions is examination of films of peripheral blood. Because of the subjective variability in blood film examination, objective measures of these changes were sought. Early in this century, Price-Jones (1910, 1922, 1933) published results of his painstaking measurements of red cell diameters. These measurements of cell diameter

Address for reprint requests: Christine E. McLaren, Ph.D., Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, Ohio 44106.

Key words: EM algorithm; Grouped doubly-truncated lognormal data; Maximum likelihood; Red blood cell volume distributions.

provided an indirect indicator of a cell's size (volume) since thickness was not determined. He plotted the frequency distribution of red cell diameters and noted a resemblance to the Gaussian distribution in healthy persons and cases of anemia following hemorrhage but a deviation from the Gaussian distribution in cases of pernicious anemia. From this he hypothesized that the differences in the distribution of red cell diameter might be diagnostically useful. Daland, Heath, and Minot (1946) found that red blood cell diameter distributions in pernicious anemia differed from those in certain other macrocytic anemias. However, the technical demands of this laborious and time-consuming procedure greatly limited its diagnostic application and the method was never widely used.

Manual techniques for the determination of mean cell volume were developed but automated methods have proven to make much more precise measurements (Wintrobe et al., 1981). Recent models of electronic particle counters not only determine the mean red cell volume but also measure the volume of each red cell, providing the red cell volume distribution. In clinical hematology laboratories, studies to detect and classify anemias are now performed almost universally by electronic particle counters, using either the aperture impedance method or darkfield illumination techniques that measure the diffraction or scattering of light by cell suspensions (Wintrobe et al., 1981).

Although previous investigators have studied the distributions of red cell volumes to determine characteristics specific to healthy individuals, there is no accepted analytic model for red cell volume. Erythrocyte distributions have been described as skewed unimodal (Mattern, Brackett, and Olson, 1957; Grant, Britten, and Kurtz, 1960; Brecher et al., 1962; Barr and Zeitler, 1964; Weed and Bowdler, 1967; England and Down, 1974) and symmetric unimodal (Bessman and Johnson, 1975; Bessman, 1980a, 1980b). Despite lack of agreement about the form of the distribution, a number of summary statistics have been suggested for the analysis of erythrocyte volume distributions including various cell volume ratios or ranges (Bessman and Johnson, 1975; England and Down, 1974) and the mean, standard deviation, and coefficient of variation of the distribution (Bessman and Feinstein, 1979). As a part of routine blood analysis, the Coulter S-Plus series, one of the most widely used types of aperture impedance particle counters, reports the arithmetic mean cell volume (MCV) as a measure of central tendency and the red cell volume distribution width (RDW) as a measure of dispersion. The RDW approximates the coefficient of variation of the distribution (Coulter Electronics, 1980). However, parameters derived from the red cell volume distribution, such as the MCV and RDW, are inherently limited because they cannot reliably summarize all the potentially useful information contained in abnormal distributions. In particular, recognizing changes in the form of the red cell volume distribution may be diagnostically valuable because some causes of anemia leave the shape unaltered (e.g., many hypoproliferative states), while in other disorders the form is changed, sometimes by the presence of subpopulations of red cells (e.g., sideroblastic anemia, developing nutritional anemia). Other investigators have provided evidence that cell volume studies can be clinically useful for the diagnostic evaluation of hematologic abnormalities, for screening patients for blood disorders, and for monitoring responses to therapy or progression of an underlying disease (Bessman, 1977, 1980a, 1980b; Bessman and Feinstein, 1979; Bessman and Johnson, 1975; Daland et al., 1946; England and Down, 1974; Johnson, Tegos, and Beutler, 1983; Proctor, Cox, and Sheridan, 1976).

Recognizing the need for uniform methodology, the International Committee for Standardization in Haematology (ICSH) has recently proposed a reference method for the analysis of cell volume distribution curves (ICSH, 1982). In brief, the ICSH recommends that an observed cell volume distribution should be fitted to one or more theoretical (reference) distributions, goodness of fit assessed, and new distributions fitted iteratively until the fit to a single distribution or to a sum of distributions is judged adequate. No methods for assessing goodness of fit or for estimating reference distribution parameters were suggested. At present, only empirical techniques for estimating fit are available because,

in part, previous statistical tests have consistently found apparently significant differences between observed and reference distributions (England, 1982). However, cell volume distributions obtained by electronic particle counters are an example of "incomplete" data (Rubin, 1976); the distributions are typically reported in grouped form (as frequency counts for given cell volume intervals) and are truncated to eliminate artifactual frequency counts. The expectation–maximization method is an iterative procedure that may be applied to "incomplete" data to provide estimates of cell volume distribution parameters (Dempster, Laird, and Rubin, 1977).

This paper describes the theoretical basis for the statistical evaluation of red blood cell volume distributions. In accordance with the recommendations of the ICSH, we have selected a reference distribution for fitting to red cell volume data. The reference distribution chosen was that predicted by a model (Section 2) of the form of the distribution of red cell volumes resulting from normal erythropoiesis. We then devised specific techniques for truncation of red cell volume data obtained by a widely used aperture impedance counter (Coulter Model S-Plus II), for iterative parameter estimation using expectation–maximization techniques, and for evaluation of goodness of fit to the reference distribution. The application of these statistical procedures to data obtained from healthy volunteers and anemic patients is illustrated.

2. Mathematical Model

The physiology of red cell production suggests that a model of the distribution of the volumes of circulating red cells may be developed. In the healthy adult, red cells are produced in the bone marrow. This process, termed erythropoiesis, consists of an orderly and irreversible process of multiplication and maturation beginning with stem cells that undergo division to maintain a population of stem cells and to produce pronormoblasts. For healthy individuals, the terminology describing the stages of erythroid maturation and the notation adopted here for the volume at each stage are listed in Table 1 and described more fully below. This simplified representation of the complex process of erythropoiesis may be adequate for the purpose of characterizing the form of the distribution of red cell volumes in healthy individuals, but could not be applied to individuals with disorders of red cell production without a much more detailed consideration of the factors regulating red cell volume (Finch, 1982). The underlying assumption of the model of the distribution of red cell volumes developed here is that, for small particles, lognormal distributions are produced if the variation in size at each step of the formation of the particle is a random proportion of the size reached at the preceding step (Aitchison and Brown, 1957, pp. 22–23).

The pronormoblast (alias the proerythroblast) is a large cell of about 800 fl that normally matures over a 4- to 6-day period while undergoing three mitotic divisions (Wintrobe et

Table 1

Mathematical notation representing the volume X of marrow and peripheral blood cells at each stage i of erythroid maturation. Let X_0 be the initial volume of the pronormoblast and $X_0/2$ be the volume after cell division.

Stage	Erythroid cell	Initial volume	Volume after growth or size reduction
1	Basophilic normoblast	$X_0/2$	X_1
2	Polychromatophilic normoblast	$X_1/2$	X_2
3	Orthochromatic normoblast	$X_2/2$	X_3
4	Marrow reticulocyte	X_3/p	X_4
5	Peripheral blood reticulocyte	X_4	X_5
6	Adult red cell	X_5	X_6

al., 1981). From the pronormoblast stage until the nucleus of the red cell is extruded, erythroid maturation can be conceived of as a sequence of n steps of volume change with hemoglobin synthesis interspersed with mitotic divisions (Table 1). This process can be expressed mathematically. After cell division of the pronormoblast, let $k = 1, \ldots, n_i$ be small steps of cell growth (stages $i = 1, 2, 3$), loss of reticulum and mitochondria (stage $i = 4, 5$), or loss of volume from aging (stage $i = 6$). Under the hypothesis that the volume variation is, at each step k, a random proportion of the cell volume X reached at the preceding step $k - 1$ (Paulus, 1975),

$$(X_k - X_{k-1})/X_{k-1} = v_{ik}, \tag{1}$$

where the v_{ik} are random and mutually independent. If the volume variation, $X_k - X_{k-1}$, at each step is small, then $dx = X_k - X_{k-1}$ and

$$\begin{aligned} \sum_{k=1}^{n_i} v_{ik} = \sum_{k=1}^{n_i} \frac{X_k - X_{k-1}}{X_{k-1}} &= \int_{X_{i-1}/2}^{X_i} \frac{dx}{X} \\ &= \log(X_i) - \log(X_{i-1}/2). \end{aligned} \tag{2}$$

Thus,

$$\log(X_i) = \log(X_{i-1}/2) + \sum_{k=1}^{n_i} v_{ik}. \tag{3}$$

By recursion, from the basophilic normoblast stage ($i = 1$) to the orthochromatic normoblast stage ($i = 3$),

$$\log(X_3) = \log(X_0/2^3) + \sum_{i=1}^{3} \sum_{k=1}^{n_i} v_{ik}. \tag{4}$$

Although on average cell division occurs 3 times, division may occur normally 2 to 4 times. If m is the average number of cell divisions from the pronormoblast stage to the cessation of cell division then

$$\log(X_3) = \log(X_0/2^m) + \sum_{i=1}^{3} \sum_{k=1}^{n_i} v_{ik}. \tag{5}$$

Equation (5) can be expanded to take into account volume variation from size reduction. By the orthochromatic normoblast stage ($i = 3$), the cell can no longer synthesize deoxyribonucleic acid and division stops. After the nucleus of the cell is extruded, the cell is classified as a reticulocyte ($i = 4$). During the next 1 to 2 days, the reticulocyte remains in the marrow, progressively loses reticulum (RNA) and mitochondria, and shrinks in size. At this stage the cell is still slightly larger than an adult red cell. After release into the bloodstream, the cell is sequestered in the spleen and "remodelled" over a period of about a day (Song and Groom, 1972), losing both cell membrane and volume and assuming a biconcave discoidal form. Over the next 1 to 2 days, as a peripheral blood reticulocyte ($i = 5$), maturation to the adult form is completed. The adult red cell ($i = 6$), now with a volume of about 90 fl, has a lifespan of 74 to 153 days (Cavill et al., 1977). During this time in the circulation, slight progressive losses of cell membrane occur with corresponding increases in density. Senescent red cells are probably selectively destroyed by reticuloendothelial cells, predominantly in the spleen in healthy individuals (Wintrobe et al., 1981).

If $1/p$ is the proportion of the cell volume remaining after the nucleus is extruded and at each stage i, $i = 4, 5, 6$, the volume variation from loss of reticulum and mitochondria and from aging at each step k is a random proportion of the volume at the preceding step

$k - 1$, then by recursion

$$\begin{aligned} \log(X_6) &= \log(X_3/p) + \sum_{i=4}^{6} \sum_{k=1}^{n_i} v_{ik} \\ &= \log(X_0/2^m p) + \sum_{i=1}^{6} \sum_{k=1}^{n_i} v_{ik}. \end{aligned} \tag{6}$$

From the central limit theorem, $\log(X_6)$ is approximately normally distributed with mean

$$\mu = \log(X_0/2^m p) + \sum_{i=1}^{6} \sum_{k=1}^{n_i} \mu_{v_{ik}} \tag{7}$$

and variance

$$\sigma^2 = \sum_{i=1}^{6} \sum_{k=1}^{n_i} \sigma^2_{v_{ik}}. \tag{8}$$

3. Methods

3.1 *Red Blood Cell Volume Distributions*

A 5.0-ml blood sample was collected from each of 50 healthy individuals and from patients with anemia. Red blood cell volume determinations were made for each sample using a Coulter particle counter Model S-Plus II. The electronic processing of aperture impedance data to correct for nonaxial cell flow, cell coincidence and aggregation, and other factors before sorting pulses into volume intervals has been described elsewhere (Jones, 1982; Rowan, 1983). For each sample frequency counts for 256 cell volume intervals of length 1.3125 fl with range 24 to 360 fl were measured in triplicate. In the upper ranges of cell volume, particle counts resulting from cell coincidence, doublets, triplets, and agglutinated cells, and in the lower ranges, particle counts from platelet clumps, large platelets, and electrical interference result in artifactual frequency counts in these ranges (Coulter Electronics, 1980). To eliminate artifactual frequency counts found in both tails of the distributions, a truncation algorithm that could be applied uniformly to each distribution was developed specifically for data collected from the Coulter Counter Model S-Plus series. For each sample the average of frequency counts for the three apertures in each cell volume interval (channel) was computed. Use of the average frequency count gives identical truncation limits for each of the apertures. The minimum cell volume and maximum cell volume having an average frequency count greater than or equal to 30 were recorded, and for each of the three apertures, all cell volume intervals between these minimum and maximum cell volume intervals were used for distribution analysis. Although the term "truncated" is sometimes used in a restricted sense to describe data that have been cut off by failure of a measuring device to record either the actual values or the number of sample points in certain ranges (Dempster et al., 1977), we use the term also to describe the data remaining after application of the algorithm to cut off data known to contain artifactual frequency counts.

3.2 *Parameter Estimates, Iterative Fitting Procedure, and Goodness-of-Fit Test*

Given x such that $\log(x)$ is from a normal distribution with mean μ and variance σ^2, then x is said to be from a lognormal distribution. To test for goodness of fit to the truncated lognormal distribution, a distribution fitting method (Hasselblad, Stead, and Galke, 1980) was modified to apply to doubly-truncated data. The maximum likelihood parameter

estimates were calculated using the expectation–maximization (EM) algorithm as described by Dempster et al. (1977). The doubly-truncated lognormal distribution can be written as

$$f(x) = \begin{cases} \dfrac{\exp\{-[\log(x)-\mu]^2/(2\sigma^2)\}}{[x\sigma(2\pi)^{1/2}]}\{\Phi[\log(T_R)] - \Phi[\log(T_L)]\}^{-1}, & \text{for } T_L < x < T_R; \\ 0, \quad \text{otherwise;} \end{cases} \tag{9}$$

where $\Phi(x)$ is the cumulative normal probability integral with mean μ and variance σ^2.

Initial estimates of μ and σ^2 were calculated using maximum likelihood formulas assuming that the observed data within a cell volume interval i fell at the midpoint. The natural log of cell volume was used for parameter estimation. These estimates can be written as

$$\begin{aligned} \mu &= \sum_{i=1}^{m} F_i \log(x_i)/N, \\ \sigma &= \left\{\sum_{i=1}^{m} F_i[\log(x_i) - \mu]^2/N\right\}^{1/2}, \\ N &= \sum_{i=1}^{m} F_i, \end{aligned} \tag{10}$$

where F_i is the frequency count in interval i, m is the total number of intervals, and N is the sample size. The E step of the expectation–maximization algorithm consists of using the initial maximum likelihood estimates of μ and σ^2 to estimate the following quantities: (i) the expected number of unobserved samples, N_L, from the left portion of the distribution; (ii) the expected number, N_R, from the right portion; (iii) the expected value of $z = \log(x)$ in each interval [$E(z_i \mid z_i$ is in a given interval)]; and (iv) the expected value of z_i^2 [$E(z_i^2 \mid z_i$ is in a given interval)]. N_L and N_R are estimated by

$$\hat{N}_L = N\Phi[\log(T_L)]/\{\Phi[\log(T_R)] - \Phi[\log(T_L)]\}, \tag{11}$$

$$\hat{N}_R = N\{1 - \Phi[\log(T_R)]\}/\{\Phi[\log(T_R)] - \Phi[\log(T_L)]\}. \tag{12}$$

The expressions for $E(z_i \mid z_i$ is in a given interval) and $E(z_i^2 \mid z_i$ is in a given interval) are given by Hasselblad et al. (1980).

The M step of the expectation–maximization algorithm then computes the estimates of μ and σ from the sufficient statistics:

$$\begin{aligned} \hat{\mu} = \Bigg\{\sum_{i=1}^{m} F_i E(z_i \mid z_i \text{ in } i\text{th interval}) &+ \hat{N}_L E[z \mid z < \log(T_L)] \\ &+ \hat{N}_R E[z \mid z > \log(T_R)]\Bigg\} \\ &\cdot\{\Phi[\log(T_R)] - \Phi[\log(T_L)]\}/N; \end{aligned} \tag{13}$$

$$\begin{aligned} \hat{\sigma}^2 = \Bigg\{\sum_{i=1}^{m} F_i E[(z_i - \mu)^2 \mid z_i \text{ in } i\text{th interval}] &+ \hat{N}_L E[(z-\mu)^2 \mid z < \log(T_L)] \\ &+ \hat{N}_R E[(z-\mu)^2 \mid z > \log(T_R)]\Bigg\} \\ &\cdot\{\Phi[\log(T_R)] - \Phi[\log(T_L)]\}/N \end{aligned} \tag{14}$$

The estimates computed in the M step are then used in the E step and this process is repeated until the desired accuracy is reached.

The chi-square goodness-of-fit test was used to test the hypothesis that x came from a doubly-truncated lognormal distribution. The χ^2 statistic with ν degrees of freedom,

$$\chi^2_\nu = \sum_{i=1}^{m} (F_i - E_i)^2/E_i, \tag{15}$$

has an approximate χ^2 distribution with $\nu = (m - 1 - p)$ degrees of freedom, where m is the number of categories and p is the number of independent parameters estimated from the data. Using the final parameter estimates, we obtained theoretical expected frequency counts E_i from the formula

$$E_i = \frac{N|\Phi(z_i) - \Phi[\log(T_L)]|}{\Phi[\log(T_R)] - \Phi[\log(T_L)]}. \tag{16}$$

Observed and expected frequency counts were compared for each of the m intervals between the left and right truncation points where the first interval was

(left truncation point, left truncation point + 1.3125]

and the last interval was

(right truncation point − 1.3125, right truncation point].

3.3 *Aperture Selection*

To provide uniform limits for truncation, the algorithm discussed in Section 3.1 was based on the average of the number of red cells for each of three apertures in cell volume intervals. The chi-square goodness-of-fit test, however, assumes an integer frequency count in each interval. Thus, to satisfy this assumption and to provide single evaluations of the goodness of fit to the doubly-truncated lognormal distribution for each independent sample, data from a single aperture were used.

The Coulter Model S-Plus II uses a system with three apertures to avoid data errors from obstructions that may affect a single aperture. The manufacturer compares data from the three apertures using a process called "voting." If the data from an aperture are judged to be outside "an established statistical range," the data from that aperture are "voted out" and the average of the data from the remaining apertures is used in further calculations (Coulter, 1980). As the procedure for determining the "established statistical range" was not specificed, we developed an alternative statistical method to lessen or eliminate the influence of extreme values caused by obstructions or other artifacts affecting a single aperture. To select an aperture to be used for analysis, the frequency distribution of the single aperture with the minimum sum of squared deviations from the average count was used for goodness-of-fit testing.

4. Results

The entire distribution from a representative sample is shown in Figure 1A. Values plotted on the vertical axis are the number of red cells sized in cell volume intervals for the single particle counter aperture used for parameter estimation. The corresponding distribution from the truncated sample is shown in Figure 1B.

Table 2 shows the truncated distribution and goodness-of-fit testing for this sample along with the expected counts and chi-square values for fit to the doubly-truncated lognormal model. For all samples probability plots (Bickel and Doksum, 1977, pp. 381–387) of the truncated data with log cell volume on the horizontal axis and corresponding normal

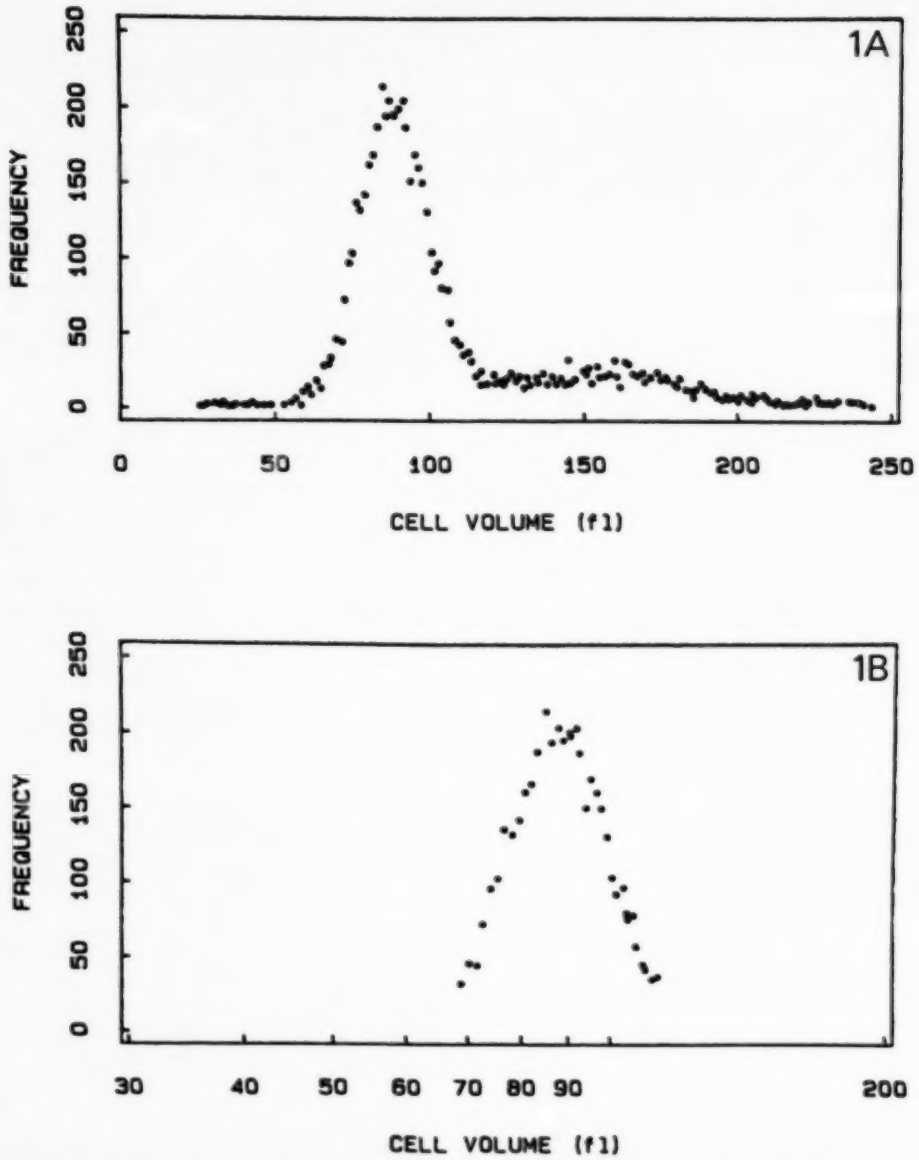


Figure 1. Frequency distribution of red blood cell volume from sample number 28 plotted on a linear scale (1A) and doubly-truncated frequency distribution plotted on a logarithmic scale (1B).

quantiles on the vertical axis were approximately linear throughout, providing evidence that the underlying distribution was lognormal. Each of the truncated empirical distributions was tested for goodness of fit to the truncated lognormal distribution. The results are shown in Table 3.

Use of the truncated lognormal model with EM parameter estimates resulted in a fairly even distribution of P-values: $.01 \leq P < .10$, $n = 9$; $.10 \leq P < .20$, $n = 8$; $.20 \leq P < .30$,

Table 2
Doubly-truncated red blood cell volume distribution from sample number 28

Cell volume interval	Observed count	Expected count	Chi-square
67.3125 to 68.6250	32	33.93	.11
68.6250 to 69.9375	45	44.46	.01
69.9375 to 71.2500	44	56.74	2.86
71.2500 to 72.5625	72	70.61	.03
72.5625 to 73.8750	97	85.78	1.47
73.8750 to 75.1875	103	101.88	.01
75.1875 to 76.5000	136	118.41	2.61
76.5000 to 77.8125	132	134.82	.06
77.8125 to 79.1250	142	150.53	.48
79.1250 to 80.4375	162	164.97	.05
80.4375 to 81.7500	167	177.62	.63
81.7500 to 83.0625	188	188.01	.00
83.0625 to 84.3750	215	195.84	1.87
84.3750 to 85.6875	194	200.86	.23
85.6875 to 87.0000	204	203.01	.00
87.0000 to 88.3125	195	202.32	.26
88.3125 to 89.6250	199	198.93	.00
89.6250 to 90.9375	204	193.11	.61
90.9375 to 92.2500	187	185.16	.02
92.2500 to 93.5625	150	175.46	3.69
93.5625 to 94.8750	169	164.42	.13
94.8750 to 96.1875	160	152.40	.38
96.1875 to 97.5000	150	139.82	.74
97.5000 to 98.8125	131	127.01	.13
98.8125 to 100.1250	103	114.29	1.12
100.1250 to 101.4375	91	101.90	1.17
101.4375 to 102.7500	96	90.07	.39
102.7500 to 104.0625	79	78.95	.00
104.0625 to 105.3750	78	68.64	1.28
105.3750 to 106.6875	57	59.23	.08
106.6875 to 108.0000	45	50.72	.65
108.0000 to 109.3125	41	43.13	.10
109.3125 to 110.6250	35	36.42	.06
110.6250 to 111.9375	37	30.55	1.36
Chi-square = 22.60, df = 31, $P = .8632$			

$n = 9$; $.30 \leq P < .40$, $n = 3$; $.40 \leq P < .50$, $n = 5$; $.50 \leq P < .60$, $n = 4$; $.60 \leq P < .70$; $n = 3$; $.70 \leq P < .80$, $n = 4$; $.80 \leq P < .90$, $n = 2$; $.90 \leq P < 1.00$, $n = 3$. Based on the one-sample chi-square test, this distribution of independent P-values was not significantly different from U(0, 1) ($\chi^2 = 12.8$, df = 9, $P = .172$). For all distributions tested, the sample sizes for individual distributions were large (3000 to 5000 cells), presumably providing adequate power for the test of goodness of fit to the lognormal model. For each truncated sample, 95–99% of the red cells were contained between the limits of the doubly-truncated distribution.

These statistical procedures were used to analyze distributions of erythrocyte volumes from patients with anemia. Figure 2A shows the red cell volume distribution from a patient with iron deficiency anemia and Figures 2B–2D show the response to iron treatment. Initially (Figure 2A) the mean was decreased ($\hat{\mu} = 4.31$) and the variation in cell volume was increased ($\hat{\sigma} = .216$) when compared to the range of $\hat{\mu}$ and $\hat{\sigma}$ found in controls ($4.38 \leq \hat{\mu} \leq 4.56$, $.118 \leq \hat{\sigma} \leq .140$). The fit to the lognormal model was good ($\chi^2 = 42.5$, df = 41, $P = .408$), providing evidence for a single population of cells. One week after treatment

Table 3
EM *parameter estimates, chi-square statistics with degrees of freedom df, and P-values for goodness of fit to the truncated lognormal distribution, for distributions sampled from 50 healthy individuals*

Sample	N	$\hat{\mu}$	$\hat{\sigma}$	χ^2	df	P
1	4123	4.51	.128	38.43	34	.2759
2	4702	4.41	.134	40.71	33	.1673
3	4121	4.41	.119	42.83	28	.0362
4	4321	4.45	.133	10.04	31	.9999
5	4148	4.46	.122	36.11	31	.2418
6	4505	4.46	.124	33.57	32	.3910
7	3986	4.49	.123	34.39	31	.3088
8	4434	4.44	.126	25.25	30	.7127
9	4459	4.42	.129	50.02	32	.0222
10	3921	4.48	.129	32.74	32	.4307
11	4025	4.45	.128	36.61	31	.2244
12	3925	4.53	.123	29.59	32	.5889
13	3968	4.49	.134	43.49	32	.0846
14	3970	4.52	.129	27.24	33	.7490
15	4311	4.48	.125	38.66	32	.1941
16	4358	4.47	.119	28.40	29	.4968
17	4290	4.47	.121	39.42	30	.1166
18	4375	4.49	.123	22.07	32	.9055
19	3987	4.44	.121	47.71	31	.0280
20	4495	4.38	.136	27.34	31	.6550
21	4256	4.52	.121	33.16	34	.5088
22	4197	4.42	.127	27.12	30	.6172
23	4411	4.46	.129	36.59	32	.2641
24	4285	4.42	.122	34.88	29	.2086
25	4307	4.49	.122	30.73	31	.4799
26	4254	4.41	.124	34.57	30	.2587
27	4215	4.55	.133	44.43	33	.0884
28	4140	4.48	.128	22.60	31	.8632
29	4333	4.46	.129	39.66	32	.1656
30	3929	4.51	.120	23.48	30	.7949
31	4186	4.40	.118	26.43	27	.4951
32	4046	4.56	.124	35.82	32	.2940
33	4493	4.43	.133	33.22	32	.4076
34	4716	4.44	.132	45.80	34	.0852
35	4375	4.46	.123	47.27	33	.0513
36	4503	4.38	.133	36.86	32	.2541
37	4865	4.50	.123	34.08	32	.3677
38	4132	4.50	.119	20.87	29	.8639
39	4082	4.48	.122	39.41	31	.1429
40	4134	4.44	.123	24.75	30	.7372
41	4233	4.45	.132	29.79	32	.5788
42	4120	4.48	.128	20.97	31	.9127
43	4707	4.38	.132	49.62	31	.0183
44	4282	4.42	.140	40.57	33	.1713
45	3910	4.45	.123	33.98	29	.2399
46	3747	4.48	.126	28.18	29	.5084
47	3964	4.51	.123	48.45	31	.0238
48	4086	4.45	.118	38.10	29	.1202
49	4379	4.51	.122	42.32	33	.1284
50	4695	4.47	.139	31.12	34	.6093

(Figure 2B) a new subpopulation of cells, now of normal size, has emerged and is reflected in the poor fit to the lognormal model ($\chi^2 = 203.53$, df = 58, $P < .0001$). Twelve weeks after treatment (Figure 2C), the new subpopulation of normally sized cells has increased and the fit to the lognormal model remains poor ($\chi^2 = 393.80$, df = 47, $P < .0001$) because

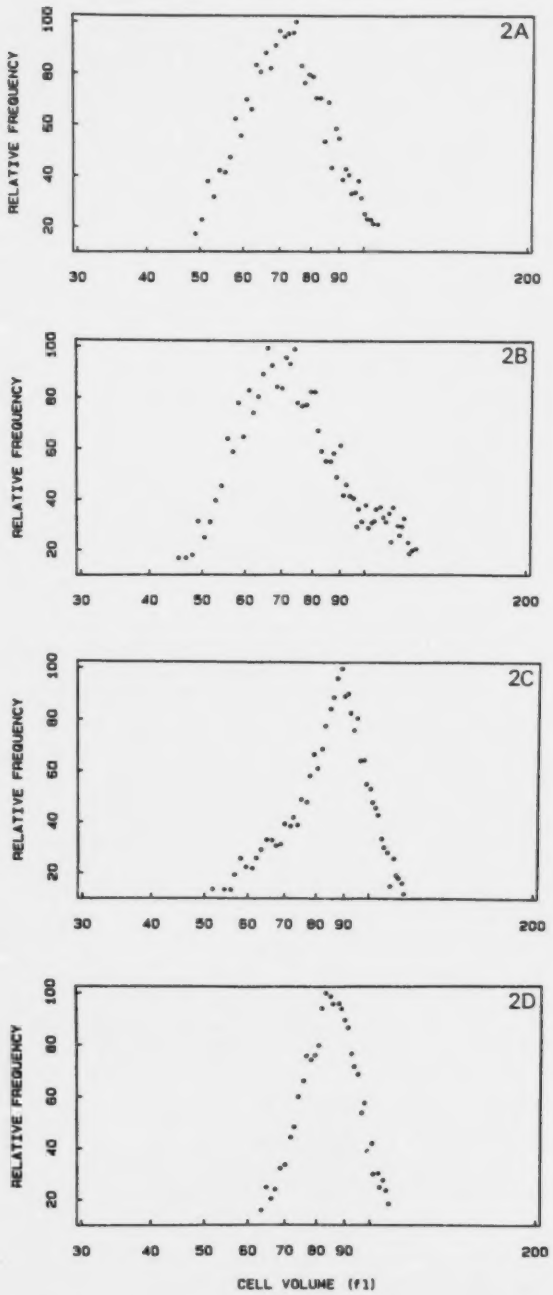


Figure 2. Frequency distributions of red blood cell volume for a patient with iron deficiency anemia: **(2A)** initial sample, $\hat{\mu} = 4.31$, $\hat{\sigma} = .216$, $\chi^2 = 42.47$, df = 41, $P = .408$; **(2B)** 1 week after treatment, $\hat{\mu} = 4.35$, $\hat{\sigma} = .293$, $\chi^2 = 203.53$, df = 58, $P < .0001$; **(2C)** 12 weeks after treatment, $\hat{\mu} = 4.46$, $\hat{\sigma} = .207$, $\chi^2 = 393.80$, df = 47, $P < .0001$; **(2D)** 16 weeks after treatment, $\hat{\mu} = 4.44$, $\hat{\sigma} = .139$, $\chi^2 = 43.25$, df = 32, $P = .089$. The relative frequency (frequency/maximum frequency × 100) of cell volume is plotted on a logarithmic scale.

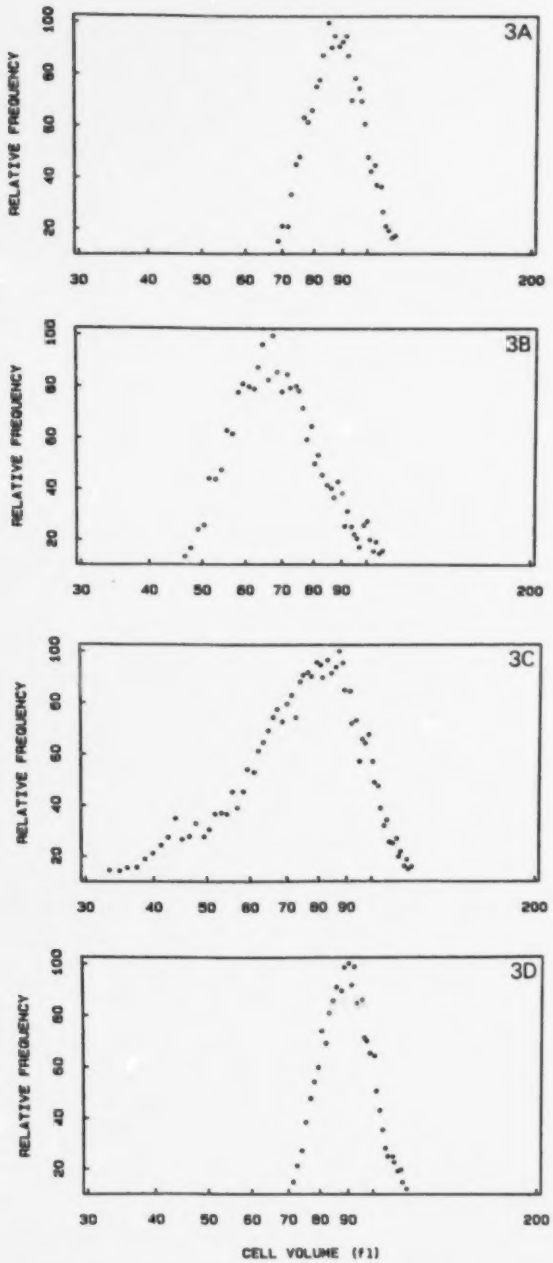


Figure 3. Frequency distributions of red blood cell volume from (**3A**) a healthy individual, $\hat{\mu} = 4.48$, $\hat{\sigma} = .128$, $\chi^2 = 22.60$, df = 31, $P = .863$; (**3B**) a patient with sideroblastic anemia $\hat{\mu} = 4.25$, $\hat{\sigma} = .220$, $\chi^2 = 102.72$, df = 44, $P < .0001$; (**3C**) a patient with sideroblastic anemia, $\hat{\mu} = 4.38$, $\hat{\sigma} = .328$, $\chi^2 = 521.98$, df = 63, $P < .0001$; and (**3D**) a patient with chronic renal failure, $\hat{\mu} = 4.49$, $\hat{\sigma} = .120$, $\chi^2 = 30.71$, df = 32, $P = .532$. The relative frequency (frequency/maximum frequency × 100) of cell volume is plotted on a logarithmic scale.

of the persistence of a subpopulation of abnormal small erythrocytes. Sixteen weeks after treatment (Figure 2D), a single population of normally sized cells remains ($\chi^2 = 43.25$, df = 32, $P = .089$) and the mean ($\hat{\mu} = 4.44$) and standard deviation ($\hat{\sigma} = .139$) were within the ranges for $\hat{\mu}$ and $\hat{\sigma}$ found in healthy individuals.

The red blood cell volume distribution of a healthy individual (sample number 28) is shown in Figure 3A. Figures 3B–3C show the distributions of two patients with sideroblastic anemia, both with morphological evidence of subpopulations of abnormal small red cells. Analyses of the truncated distributions showed that the hypothesis of lognormality was rejected for both. Figure 3D shows a volume distribution from a patient with chronic renal failure. The fit to the lognormal is good ($\chi^2 = 30.71$, df = 32, $P = .532$), and the mean and standard deviation are within the ranges found in healthy individuals ($\hat{\mu} = 4.49$, $\hat{\sigma} = .120$).

5. Discussion

Various measures of central tendency and variability have been proposed for the distribution of red blood cell volume in healthy individuals, although without having first established the fit to an underlying distribution. Our results demonstrate that difficulties encountered by previous investigators in statistical analysis of these distributions may be ascribed in part to (i) uncertainty about the underlying family of distributions; (ii) lack of a data-truncation method for the elimination of frequency counts from cell coincidence, agglutinated cells, platelets, and debris; and (iii) lack of an appropriate iterative fitting procedure and parameter estimation method for incomplete data. Instead of testing empirical data against the many possible theoretical families of distributions, we first studied the physiology of normal red cell production and formulated a mathematical model which takes into account volume variation from cell growth, volume loss, nuclear division, and aging, and predicts a lognormal distribution of red cell volumes. This model provided a theoretical basis for testing goodness of fit to a particular family of distributions. Previous studies of red cell volume distributions have shown that particle counts resulting from cell coincidence, agglutinated cells, platelets, and debris should be eliminated (Grant et al., 1960; Bull, 1968; England, 1982). An examination of representative distributions from healthy individuals showed fluctuation in both tails of the distributions. We then developed a truncation algorithm specific to data collection from the Coulter Counter Model S-Plus series that could be applied uniformly to each distribution to eliminate artifactual frequency counts found in the extreme tails. Because the truncated empirical data were not continuous but consisted of grouped frequency counts, it was necessary to apply a method for lognormal parameter estimation appropriate to grouped truncated distributions. The problem of parameter estimation from incomplete data was solved using a technique known as the missing information principle or expectation–maximization algorithm (Orchard and Woodbury, 1972; Dempster et al., 1977). The one-sample chi-square test was used to test for goodness of fit to the doubly-truncated lognormal distribution. If the model is correct P-values for goodness of fit should be distributed independently as U(0, 1). Our results show that the distribution of P-values from fifty independent samples was not significantly different from the U(0, 1) distribution, providing evidence that the truncated lognormal distribution is suitable for use as a reference distribution for goodness-of-fit testing.

The limitations of our analysis should be acknowledged. For example, our procedure does not recognize that each of the F_i in the cell volume intervals is an estimated count measured with error. Although this additional complexity might be handled within the EM framework, we have not refined the E-step expectations to adjust for the counting error. In addition, our results may also have been influenced by the numbers of cells counted, the interval size and number of intervals, or by other factors affecting the measurement of cell volumes by aperture impedance (Rowan, 1983).

Examples of the application of these procedures to samples obtained from patients with anemia are shown in Figures 2 and 3. Lognormal distributions were identified in a healthy individual (Figure 3A), in a patient with chronic iron deficiency anemia, initially (Figure 2A), and after completion of therapy for iron deficiency (Figure 2D), and in a patient with the anemia associated with chronic renal failure (Figure 3D). By contrast, the emergence of subpopulations of cells during the treatment of iron deficiency (Figures 2B and 2C) produced non-lognormal distributions. Similarly, in the sideroblastic anemias, disorders with morphologically evident subpopulations of red cells, the red cell volume distributions are not lognormal (Figures 3B and 3C). Note that the parameters $(\hat{\mu}, \hat{\sigma})$ of the volume distributions shown in Figures 2A and 3B are similar but that the diagnostically important feature of the subpopulation of abnormally small cells in Figure 3B is indicated only after testing for lognormality. In future investigations of the distributions produced by abnormal red blood cell production, known physiologic changes could be incorporated into the model of normal erythropoiesis, providing new theoretical models for goodness-of-fit testing against empirical data found in different anemias.

We emphasize that the results presented here were obtained from analysis of frequency distributions as recorded by the Coulter S-Plus II aperture impedance particle counter and may have been affected by the techniques used for the electronic processing of aperture impedance data (Jones, 1982; Rowan, 1983) or by other features of this series of instruments. Studies of red blood cell volume distributions from other particle counters, using either aperture impedance or darkfield illumination methods, may require different analytic procedures. However, the principles of the statistical methods we describe for fitting reference distributions to incomplete data could be applied to red cell volumes determined by other instruments as well as to particles with similar distributional characteristics, such as platelets (Paulus, 1975).

The Expert Panel on Cell Counting and Sizing of the International Committee for Standardization in Haematology has recommended guidelines for analysis of cell volume distributions (ICSH, 1982). These guidelines suggest testing the goodness of fit of an observed cell volume distribution against theoretical (reference) distributions. When the fit to a specified distribution is judged to be adequate, the estimates for the parameters of that distribution should be recorded and used for further analyses. Based on our results, we suggest that the truncated lognormal distribution should be used as a reference distribution for goodness-of-fit testing for red blood cell volume distributions and that expectation–maximization parameter estimates be used for cell volume analyses.

Acknowledgements

Dr McLaren was the recipient of a Research Fellowship of the American Heart Association, Northeast Ohio Affiliate, Inc. We thank Prof. John W. Harris, Prof. Arthur Steinberg, and Dr James D. Knoke for helpful comments on the manuscript. We also thank Dr A. Richardson Jones and members of the Engineering Department of Coulter Electronics, Inc., for their technical assistance and Mrs. Lisa Pence Thum for her help with data collection and computer analysis.

Résumé

Nous avons développé des méthodes statistiques, en accord avec les recommandations générales du Comité International pour la standardisation en Hématologie (1982, *Journal of Clinical Pathology* **35**, 1320–1322), pour analyser les distributions du volume des globules rouges. Pour dériver une distribution adéquate de référence pour les tests d'adjustement, nous avons dérivé un modèle mathématique de l'erythropoièse qui prédit une forme lognormale pour la distribution des volumes des erythrocytes. On a alors testé les modèles de prédictions à l'aide d'échantillons de 50 individus sains. Chaque distribution groupée du volume des globules a été doublement tronquée pour éliminer

les artefacts. On a calculé les paramètres de la distribution un utilisant l'algorithme de maximisation de l'espérance, une technique de donnée manquante. Les résultats d'un test du chi 2 montrent une distribution plutôt plate des p-valeurs sur l'intervalle [0, 1]. On donne des exemples d'application de ces procédures statistiques sur des distributions provenant de patients atteints d'anémie. Nos résultats suggèrent que, pour l'analyse des volumes des globules rouges, (i) l'estimation des paramètres devrait être faite par la méthode de maximisation de l'espérance (algorithme EM) et (ii) la loi lognormale tronquée devrait être utilisée comme distribution de référence dans les ajustements. Cette méthode pourrait être appliquée à tout ensemble de données groupées doublement tronquées qui, après transformation, suivent un modèle normal.

References

Aitchison J. and Brown, J. A. C. (1957). *The Lognormal Distribution with Special Reference to Its Use in Economics*. Cambridge: Cambridge University Press.

Barr, G. F. and Zeitler, E. H. (1964). Letter to the editor: Distribution of erythrocytes. *Blood* **23**, 403.

Bessman, J. D. (1977). Microcytic polycythemia. Frequency of nonthalassemic causes. *Journal of the American Medical Association* **238**, 2391–2392.

Bessman, J. D. (1980a). Evaluation of automated whole-blood platelet counts and particle sizing. *American Journal of Clinical Pathology* **74**, 157–162.

Bessman, J. D. (1980b). Heterogeneity of red cell volume: Quantitation, clinical correlations, and possible mechanisms. *The Johns Hopkins Medical Journal* **146**, 226–230.

Bessman, J. D. and Feinstein, D. I. (1979). Quantitative anisocytosis as a discriminant between iron deficiency and thalassemia minor. *Blood* **53**, 288–293.

Bessman, J. D. and Johnson, R. K. (1975). Erythrocyte volume distribution in normal and abnormal subjects. *Blood* **46**, 369–379.

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco: Holden-Day.

Brecher, G., Jakobek, E. F., Schneiderman, M. A., Williams, G. Z., and Schmidt, P. J. (1962). Size distribution of erythrocytes. *Annals of the New York Academy of Science* **99**, 242–261.

Bull, B. S. (1968). On the distribution of red cell volumes. *Blood* **31**, 503–515.

Cavill, I., Ricketts, C., Napier, J. A. F., and Jacobs, A. (1977). Ferrokinetics and erythropoiesis in man: Red-cell production and destruction in normal and anaemic subjects. *British Journal of Haematology* **35**, 33–40.

Coulter Electronics, Inc. (1980). *Model S-Plus* II *Operations Manual for the Coulter Channelyzer* II. Hialeah, Florida: Coulter Electronics, Inc.

Daland, G. A., Heath, C. W., and Minot, G. R. (1946). Differentiation of pernicious anemia and certain other macrocytic anemias by the distribution of red blood cell diameters. *Blood* **1**, 67–75.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **38**, 1–22.

England, J. M. (1982). The analysis and interpretation of cell size distribution curves in hematology: A review. In *Advances in Hematological Methods: The Blood Count*, O. W. Assendelft and J. M. England (eds), 109–123. Boca Raton, Florida: CRC Press.

England, J. M. and Down, M. C. (1974). Red-cell-volume distribution curves and the measurement of anisocytosis. *The Lancet* **1**, 701–703.

Finch, C. A. (1982). Erythropoiesis, erythropoietin, and iron. *Blood* **60**, 1241–1246.

Grant, J. L., Britten, M. C., and Kurtz, T. E. (1960). Measurement of red blood cell volume with the electronic cell counter. *American Journal of Clinical Pathology* **33**, 138–143.

Hasselblad, V., Stead, A. G., and Galke, W. (1980). Analysis of coarsely grouped data from the lognormal distribution. *Journal of the American Statistical Association* **75**, 771–778.

International Committee for Standardization in Hematology. (1982). ICSH recommendations for the analysis of red cell, white cell, and platelet size distribution curves: I. General principles. *Journal of Clinical Pathology* **35**, 1320–1322.

Johnson, C. S., Tegos, C., and Beutler, E. (1983). Thalassemia minor: Routine erythrocyte measurements and differentiation from iron deficiency. *American Journal of Clinical Pathology* **80**, 31–36.

Jones, A. R. (1982). Counting and sizing of blood cells using aperture impedance systems. In *Advances in Hematological Methods: The Blood Count*, O. W. Assendelft and J. M. England (eds), 49–72. Boca Raton, Florida: CRC Press.

Mattern, C. F. T., Brackett, F. S., and Olson, B. J. (1957). Determination of number and size of particles by electrical gating: Blood cells. *Journal of Applied Physiology* **10**, 56–70.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: Theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* **1**, 697–715. Berkeley: University of California Press.
Paulus, J. (1975). Platelet size in man. *Blood* **46**, 321–336.
Price-Jones, C. (1910). The variation in sizes of red blood cells. *British Medical Journal* **2**, 1418.
Price-Jones, C. (1922). The diameters of red cells in pernicious anemia and in anemia following haemorrhage. *Journal of Pathology* **25**, 487–504.
Price-Jones, C. (1933). *Red Blood Cell Diameters*. London: Oxford University Press.
Proctor, S. J., Cox, J. R., and Sheridan, T. J. (1976). Anisocytosis and the C-1000 channelyzer in macrocytic anaemia. *Journal of Clinical Pathology* **29**, 719–723.
Rowan, R. M. (1983). *Blood Cell Volume Analysis—A New Screening Technology for the Haematologist*. London: Albert Clark and Company Limited.
Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.
Song, S. H. and Groom, A. C. (1972). Sequestration and possible maturation of reticulocytes in the normal spleen. *Canadian Journal of Physiology and Pharmacology* **50**, 400–406.
Weed, R. I. and Bowdler, A. J. (1967). The influence of hemoglobin concentration on the distribution pattern of the volume of human erythrocytes. *Blood* **29**, 297–312.
Wintrobe, M. M., Lee, G. R., Boggs, D. R., Bithell, T. C., Foerster, J., Athens, J. W., and Lukens, J. N. (1981). *Clinical Hematology*, 8th ed. Philadelphia: Lea and Febiger.

Received July 1984; *revised June and November* 1985.

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

This article documents research that was conducted under the auspices of the U.S. Environmental Protection Agency (USEPA). As an arm of the government, the USEPA has a paid up, non-exclusive, irrevocable, worldwide license to reproduce, prepare derivative works, and distribute to the public all of the material contained herein, and to grant others acting on its behalf the right to do so. The USEPA grants to the National Technical Information Service, Department of Commerce, the right to act as its agent in the distribution of this material.